



TITLE:

# Input variable scaling for statistical modeling

AUTHOR(S):

Kim, Sanghong; Kano, Manabu; Nakagawa, Hiroshi;  
Hasebe, Shinji

---

CITATION:

Kim, Sanghong ...[et al]. Input variable scaling for statistical modeling.  
Computers & Chemical Engineering 2015, 74: 59-65

ISSUE DATE:

2015-03

URL:

<http://hdl.handle.net/2433/193671>

RIGHT:

© 2015 Elsevier Ltd.; This is not the published version. Please cite only the published version.; この論文は出版社版ではありません。引用の際には出版社版をご確認ご利用ください。

# Input Variable Scaling for Statistical Modeling

Sanghong Kim<sup>a,\*</sup>, Manabu Kano<sup>b</sup>, Hiroshi Nakagawa<sup>c</sup>, Shinji Hasebe<sup>a</sup>

<sup>a</sup>*Dept. of Chemical Engineering, Kyoto University, Kyoto 6158510, Japan*

<sup>b</sup>*Department of Systems Science, Kyoto University, Kyoto, 6068501, Japan*

<sup>c</sup>*Formulation Technology Research Laboratories, Daiichi Sankyo Co., Ltd., Hiratsuka 2540014, Japan*

---

## Abstract

Input variable scaling is one of the most important steps in statistical modeling. However, it has not been actively investigated, and autoscaling is mostly used. This paper proposes two input variable scaling methods for improving the accuracy of soft sensors. One method statistically derives the input variable scaling factors; the other one uses spectroscopic data of a material whose content is estimated by the soft sensor. The proposed methods can determine the scales of the input variables based on their importance in output estimation. Thus, it can reduce the negative effects of input variables which are not related to an output variable. The effectiveness of the proposed methods was confirmed through a numerical example and industrial applications to a pharmaceutical and a distillation processes. In the industrial applications, the proposed methods improved the estimation accuracy by up to 63% compared to conventional methods such as autoscaling with input variable selection.

**Keywords:** Statistical model, Soft sensor, Input variable scaling, Pharmaceutical process, Distillation process

---

## 1. Introduction

In the process industry, one of the most important tasks is to ensure quality and to reduce operating cost. However, real-time measurement of product quality is not always available due to unacceptable measurement equipment cost and long measurement time. To solve this problem, research on soft sensors,

---

\*Corresponding author. Tel.: +81-(0)75-383-2677; fax: +81-(0)75-383-2677.

Email address: [kim@cheme.kyoto-u.ac.jp](mailto:kim@cheme.kyoto-u.ac.jp) (Sanghong Kim)

6 which estimate product quality using real-time measurements, has been actively  
7 conducted (Kadlec et al., 2009; Kano and Fujiwara, 2013; Oh et al., 2013;  
8 Khatibisepehr et al., 2014). According to a questionnaire survey (Kano and  
9 Fujiwara, 2013), in 2009 soft sensors were working in over 400 distillation and  
10 chemical reaction processes at 15 companies in Japan. In addition, soft sensors  
11 have recently attracted much interest in the pharmaceutical industry to achieve a  
12 new quality assurance system composed of Quality by Design (QbD) and process  
13 analytical technology (PAT) (Roggo et al., 2007; Rajalahti and Kvalheim, 2011).  
14 Building a soft sensor requires many steps such as data acquisition, abnormal data  
15 detection, data preprocessing, input variable selection, model building, and model  
16 validation. Although input variable scaling, a data preprocessing method in which  
17 the values of each input variable are multiplied by the scaling factor of the input  
18 variable, can have significant effect on the estimation performance of soft sensors,  
19 research on input variable scaling has not been actively conducted. Hence, this  
20 paper focuses on input variable scaling, which is mathematically represented as

$$\tilde{\mathbf{X}} = \mathbf{X}\mathbf{\Lambda} \quad (1)$$

$$\mathbf{\Lambda} = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_M) \quad (2)$$

21 where  $\mathbf{X} \in \mathbb{R}^{N \times M}$  is the raw input variable matrix, in which the input variables  
22 are not scaled,  $\tilde{\mathbf{X}} \in \mathbb{R}^{N \times M}$  is the scaled input variable matrix,  $\lambda_m$  is a nonnegative  
23 input variable scaling factor for the  $m$ -th input variable,  $N$  is the number of  
24 samples, and  $M$  is the number of input variables. It is assumed that the mean of  
25 each input variable is zero without loss of generality. The input variable scaling  
26 affects important statistical properties of the data such as the distance between  
27 samples and the covariance of samples. It also affects the estimation result.  
28 For example, the  $m$ -th input variable  $x_m$  cannot have any influence on output  
29 estimation when  $\lambda_m$  is zero. Thus,  $\mathbf{\Lambda} \in \mathbb{R}^{M \times M}$  should be carefully selected to  
30 create accurate soft sensors.

31 In past research, autoscaling was commonly used (Engel et al., 2013; van den  
32 Berg et al., 2006; Todeschini et al., 1999). In addition, Pareto scaling, level  
33 scaling, poisson scaling, range scaling, and VAST scaling (Keun et al., 2003)

34 have been considered. The scaling factors in these methods are defined as

$$\frac{1}{\lambda_m} = \begin{cases} \sigma_m & (\text{autoscaling}) \\ \sqrt{\sigma_m} & (\text{pareto scaling}) \\ \bar{x}_m & (\text{level scaling}) \\ \sqrt{\bar{x}_m} & (\text{poisson scaling}) \\ x_{m,\max} - x_{m,\min} & (\text{range scaling}) \\ \frac{\sigma_m^2}{\bar{x}_m} & (\text{VAST scaling}) \end{cases} \quad (3)$$

35 where  $\sigma_m$  is the standard deviation of  $x_m$ ,  $\bar{x}_m$  is the mean value of  $x_m$ ,  $x_{m,\max}$  is  
36 the maximum value of  $x_m$ , and  $x_{m,\min}$  is the minimum value  $x_m$ . These methods  
37 define the input variable scaling factors based only on the information from the  
38 input variables such as their standard deviations and means. Hence, input variable  
39 scaling factors can be large for the input variables which are irrelevant to the  
40 output variable when these method are used, and the estimation performance  
41 of soft sensors may deteriorate. Some of the irrelevant input variables might  
42 be removed by using input variable selection methods such as the stepwise  
43 method (Hocking, 1976), variable influence on projection (VIP) (Wold et al.,  
44 2001) and least absolute shrinkage and selection operator (LASSO) (Tibshirani,  
45 1996). It is, however, very difficult to remove all irrelevant input variables  
46 without removing any relevant input variables, and some irrelevant input variables  
47 generally remain after input variable selection. Thus, it is needed to determine the  
48 input variable scaling factors according to the importance of the input variables  
49 in output estimation. To take into account the importance of input variables  
50 in the output estimation, Kuzmanovski et al. (Kuzmanovski et al., 2009) used  
51 the genetic algorithm to optimize the input variable scaling factor. However,  
52 the computational burden of the genetic algorithm is considerable. Martens et  
53 al. (Martens et al., 2003) proposed to use the magnitude of the undesired signals  
54 in measurements to determine the input variable scaling factors. But, this method  
55 is applicable only to spectroscopic data. To solve the above-mentioned problems,  
56 two input variable scaling methods are proposed. The proposed methods can  
57 determine the input variable scaling factors based on the importance of input  
58 variables in output estimation with short computational time. One of the proposed  
59 methods can be applied to any data.

## 60 2. Input variable scaling methods

61 Conventional input variable scaling methods such as autoscaling and range  
62 scaling do not determine the input variable scaling factors based on the importance  
63 of individual input variables in output estimation. These methods, therefore, can  
64 cause overfitting especially when the number of samples is small. One can reduce  
65 the effect of irrelevant input variables on output estimation by assigning small  
66 input variable scaling factors to those input variables. On the other hand, large  
67 input variable scaling factors should be assigned to input variables which have a  
68 large influence on an output variable.

69 We propose two methods to evaluate the influence of each input variable on  
70 an output variable and assign appropriate input variable scaling factors to input  
71 variables. The first one statistically derives the input variable scaling factors, while  
72 the second one uses spectroscopic data of a material whose content is estimated  
73 by a soft sensor.

### 74 2.1. Proposed method 1: data-based approach

75 Proposed method 1 statistically calculates the input variable scaling factor in  
76 an iterative manner. In this paper, the standardized regression coefficients of input  
77 variables in a partial least squares (PLS) model and the VIP scores are used as the  
78 input variable scaling factor, since they correlate to the importance of each input  
79 variable. The standardized regression coefficient is defined as the product of the  
80 regression coefficient  $\beta$  and the standard deviation  $\sigma$  of an input variable. The  
81 algorithm of proposed method 1 is as follows:

- 82 1. Prepare the raw input variable matrix  $\mathbf{X}$  and an output variable vector  $\mathbf{y} \in \mathbb{R}^N$ .
- 83
- 84 2. Set the iteration number  $i$  to 1 and the maximum iteration number to  $I$ .
- 85 3. Calculate the input variable scaling factor matrix  $\Lambda_0 =$   
86  $\text{diag}(\lambda_{10}, \lambda_{20}, \dots, \lambda_{M0})$  where  $\lambda_{m0}$  is  $1/\sigma_{m0}$ . Here,  $\sigma_{m0}$  is the standard  
87 deviation of the  $m$ -th input variable ( $m = 1, 2, \dots, M$ ) in the raw input  
88 variable matrix  $\mathbf{X}$ .
- 89 4. Let the scaled input matrix  $\tilde{\mathbf{X}}_0 = \mathbf{X}\Lambda_0$ .
- 90 5. Calculate the new input variable scaling factor matrix

$$\Lambda_i = \text{diag}(\lambda_{1i}, \lambda_{2i}, \dots, \lambda_{Mi}) \quad (4)$$

$$\lambda_{mi} = \begin{cases} |\beta_{mi}| \sigma_{mi} & (\text{standardized regression coefficient}) \\ \text{VIP}_{mi} & (\text{VIP score}) \end{cases} \quad (5)$$

- 91 for every  $m$ . Here,  $\beta_{mi}$ ,  $\sigma_{mi}$  and  $VIP_{mi}$  denote the regression coefficient, the  
92 standard deviation and VIP score of the  $m$ -th input variable obtained using  
93 the scaled input matrix  $\tilde{\mathbf{X}}_{i-1}$  and the output variable vector  $\mathbf{y}$ , respectively.  
94 6. Calculate the new scaled input matrix  $\tilde{\mathbf{X}}_i = \mathbf{X} \Lambda_i$ .  
95 7. Finish the calculation if  $i = I$ . Otherwise set  $i = i + 1$  and go to step 5.

96 Steps 3 and 4 in the above algorithm correspond to autoscaling. In step 5, the  
97 input variable scaling factors are updated, and the input variable matrix is updated  
98 in step 6. The explicit expression of the regression coefficient in a PLS model and  
99 the VIP score is available in section 4.2 of (Kim et al., 2013). The convergence  
100 of this method is not guaranteed in all cases. However, the values of regression  
101 coefficients converged in most cases at least in the case studies conducted in this  
102 paper as shown in the next section.

103 The regression coefficient vector obtained by PLS is represented as

$$\beta_{\text{PLS}} = \mathbf{W}(\mathbf{P}^T \mathbf{W})^{-1} \mathbf{q} \quad (6)$$

$$\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_R] \quad (7)$$

$$\mathbf{P} = [\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_R] \quad (8)$$

$$\mathbf{q} = [q_1, q_2, \dots, q_R]^T \quad (9)$$

104 where  $\mathbf{w}_r$ ,  $\mathbf{p}_r$  and  $q_r$  are the weight vector, the loading vector of the input variable  
105 and the regression coefficient for the  $r$ -th latent variable.

106 The VIP score (Wold et al., 2001) of the  $m$ -th variable is defined as

$$VIP_m = \sqrt{\frac{M \sum_{r=1}^R \left[ (q_r^2 \mathbf{t}_r^T \mathbf{t}_r) \left( \frac{w_{mr}}{\|\mathbf{w}_r\|} \right)^2 \right]}{\sum_{r=1}^R (q_r^2 \mathbf{t}_r^T \mathbf{t}_r)}} \quad (10)$$

107 where  $w_{mr}$  is the  $m$ -th component of the  $r$ -th weight vector  $\mathbf{w}_r$ .  $\mathbf{t}_r$  is the  $r$ -th  
108 latent variable score.

## 109 2.2. Proposed method 2: knowledge-based approach

In the pharmaceutical and food industries, soft sensors are often used to estimate the content of an important material from the spectroscopic data of products (Cen and He, 2007; Roggo et al., 2007; Jamragiewicz, 2012). In such a situation, it is crucial to identify the important variables/wavelengths.

A large number of statistical wavelength selection methods have been proposed (Jouen-Rimbauda and Massart, 1995; Nørgaard et al., 2000; Jiang et al., 2002; Kim et al., 2011; Fujiwara et al., 2012). These methods, however, may not work well when the number of samples is small. In addition, they have tuning parameters, which are difficult to determine. To solve this problem, this paper proposes a knowledge-based input variable scaling method using the spectrum of the important material, in which the input variable scaling factor  $\lambda_m$  is defined as

$$\lambda_m = \frac{|\xi_m|}{\sigma_{x_m}} \quad (11)$$

110 where  $\xi_m$  is the (preprocessed) spectrum signal of an important material at  
111 the  $m$ -th wavelength and  $\sigma_{x_m}$  is the standard deviation of the (preprocessed)  
112 spectrum signal at the  $m$ -th wavelength in the raw input variable matrix  $\mathbf{X}$ .  
113 Here, the spectrum signals of the important material and the products might be  
114 preprocessed before the input variable scaling factor is calculated. For example,  
115 the Savitsky-Golay filter (Savitzky and Golay, 1964) and standard normal variate  
116 (SNV) (Barnes et al., 1989) can be used.

117 This method is based on the idea that the wavelengths where the ratio  
118  $\lambda_m$  is small are not important for soft-sensor design, because they have low  
119 signal-to-noise ratios and the (preprocessed) spectrum signal of the products  
120 would not significantly change with the amount of the important material at  
121 those wavelengths. Proposed method 2 is free from parameter tuning and uses  
122 process knowledge. Thus, it is expected to achieve higher estimation performance  
123 especially when the number of samples is small compared to proposed method 1,  
124 which uses only statistical information of the process data.

### 125 3. Illustrative numerical example

126 In this section, an illustrative numerical example is shown to confirm that input  
127 variable scaling can have significant influence on the estimation accuracy of soft  
128 sensors and that proposed method 1 can improve estimation accuracy.

#### 129 3.1. Problem setting

130 In this example, the number of input variables  $x_m$  is 30 and the number of  
131 output variable  $y$  is 1. Input and output variables are the sum of real values of

state variables  $s_m$  and measurement noises  $w_m$ , which are defined as follows.

$$w_m \sim N(0, 0.005^2) \quad (m = 0, 1, \dots, 30) \quad (12)$$

$$s_m \sim \text{rand}(0, 1) \quad (m = 1, 2, \dots, 30) \quad (13)$$

$$x_m = s_m + w_m \quad (14)$$

$$y = s_1 + 3s_2 + 5s_3 + w_0 \quad (15)$$

Here,  $N(\mu, \sigma^2)$  denotes the normal distribution whose mean is  $\mu$  and standard deviation is  $\sigma$ , and  $\text{rand}(a, b)$  denotes the uniform random distribution on the open interval from  $a$  to  $b$ .  $w_m$  and  $s_m$  are independent from each other.  $x_m$  and  $y$  are the measurements used for soft-sensor design while  $s_m$  and  $w_m$  are not measured.

In this example, only three input variables ( $x_1$ - $x_3$ ) are related to the output variable and the input-output relationship is linear. The other 27 variables ( $x_4$ - $x_{30}$ ), which are not related to the output variable, are used for model building. Thus, the probability of chance correlation could be high when the number of samples for model building is small. Input variable selection methods were not used to check whether input variable scaling can reduce the risk of chance correlation when irrelevant variables cannot be removed by input variable selection.

From Equations (12)-(15), 15 samples are generated and used for model building. The number of samples is realistic since it is usual that the number of samples is much smaller than that of input variables when spectroscopic data is used for soft-sensor design. For example, the number of samples for model building is 9 or 45, and the number of input variable is 1868 in the example described in Section 4.1. To validate the soft sensor built using the 15 samples, 3000 samples are independently generated and used as model validation data. It should be noted that 3000 samples are used just for model validation and not available when the soft sensor is built. In addition, because  $w_m$  and  $s_m$  are randomly determined and their values affect estimation performance, 1000 sets of model building and validation data are generated and each dataset was used separately.

For soft-sensor design, PLS was used with one of the following input variable scaling methods:

1. Autoscaling.
2. A reference method in which  $\lambda_m = 1$  ( $m = 1, 2, 3$ ) and  $\lambda_m = 0.1$  ( $m = 4, 5, \dots, 30$ ).
3. Proposed method 1 with different maximum iteration numbers  $I = 1, 3$  and 5.



164 In the reference method, larger input variable scaling factors are assigned to  
165  $x_1$ - $x_3$  than  $x_4$ - $x_{30}$ . It should be noted that the reference method cannot be  
166 used in real situations because the importance of each input variable is generally  
167 unknown. The number of the latent variables for each PLS model is determined  
168 by leave-one-out cross-validation.

### 169 3.2. Results and discussion

170 The model validation results for 1000 sets of model building and validation  
171 data are shown in Figure 1. Comparing autoscaling and the reference method  
172 confirms that the estimation accuracy can be greatly improved by properly setting  
173 the input variable scaling factors. In addition, proposed method 1 successfully  
174 reduced average of the root mean square error (RMSE) for the validation data as  
175 well as the reference method. Proposed method 1 had higher standard deviation of  
176 the RMSE than the reference method. This is because the standardized regression  
177 coefficients and the VIP scores do not always accurately represent the importance  
178 of the input variables when they are obtained from only 15 samples. Figure 2  
179 shows an example of the change of the regression coefficients for input variables  
180 before input scaling in a model building data. The values at iteration number 0  
181 are those obtained by autoscaling. The convergence is not guaranteed in all cases.  
182 However, the values of regression coefficients converged in most cases at least in  
183 the case studies conducted in this paper as shown in Figure 2.

184 In this example, smaller RMSE was obtained by using VIP scores than using  
185 the standardized regression coefficients, but the difference is not significant and  
186 using the standardized regression coefficients might be better in another example.  
187 The method for selecting the best statistical index is outside the scope of this  
188 research.

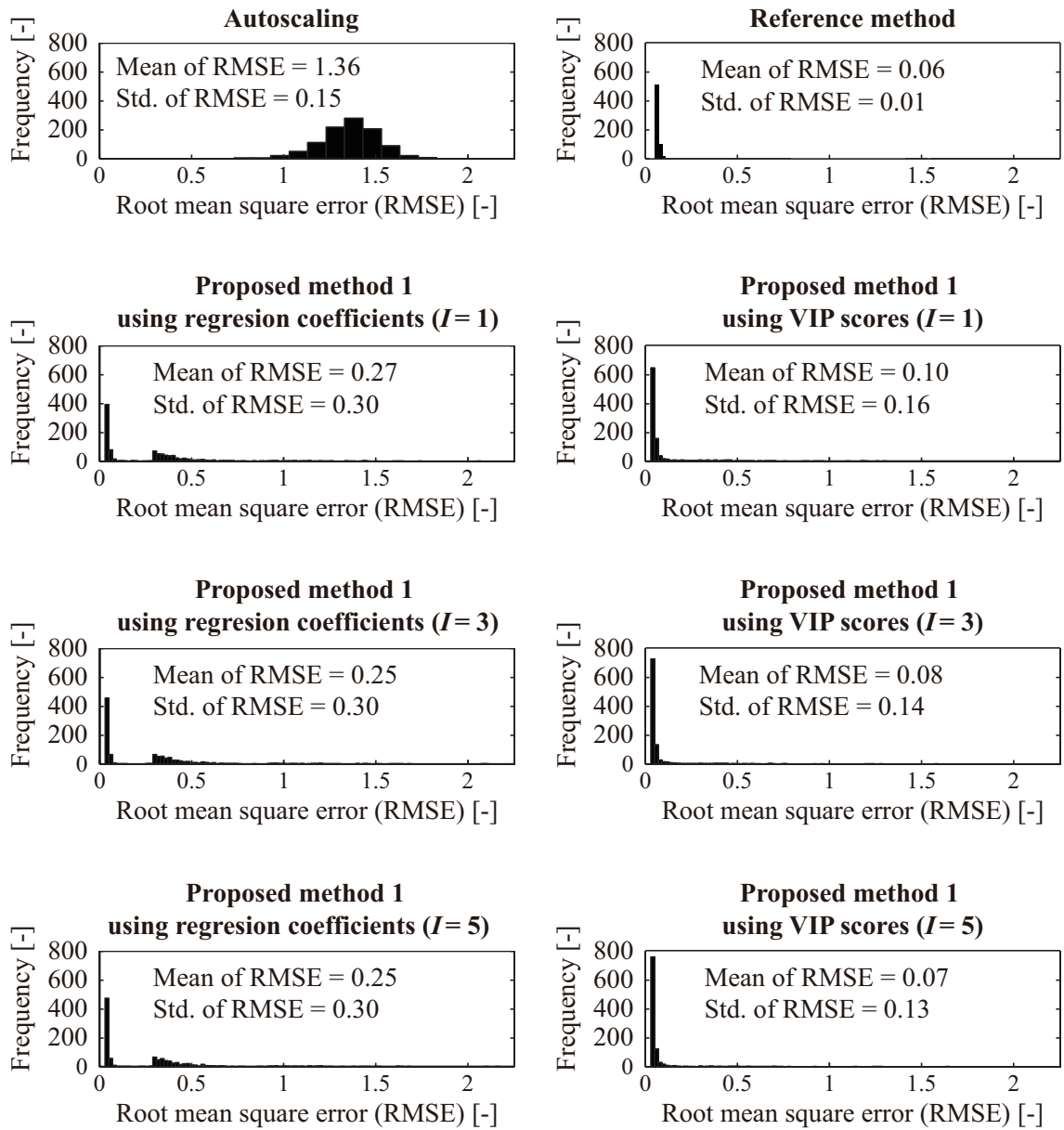


Figure 1: Model validation result for 1000 datasets in the numerical example.

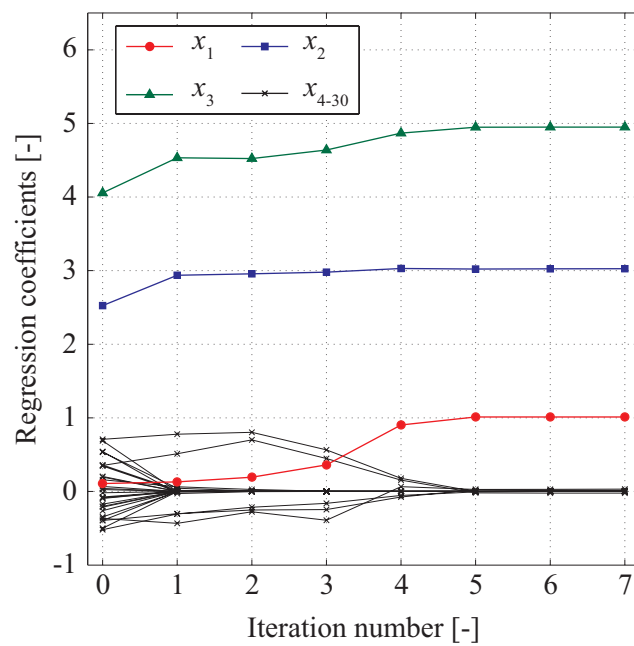


Figure 2: Change of regression coefficients for input variables before input scaling with the iteration number.

## 189 4. Industrial application

### 190 4.1. Pharmaceutical process

191 In the pharmaceutical industry, it is required to measure the amount of residual  
192 drug substances in manufacturing equipment after cleaning for product quality  
193 assurance and safety. Soft sensors are useful for achieving rapid and low-cost  
194 measurement of the amount of residual drug substances. In this paper, soft sensors  
195 were built to estimate the amount of magnesium stearate, which is a standard  
196 excipient in tablets, using the infrared spectrum of the methanol solution for  
197 different magnesium stearate concentrations. The overview of the experimental  
198 data is shown in Table 1. The absorbance spectra were measured at 400-4000  
199  $\text{cm}^{-1}$ . The spectra were secondary differentiated to reduce the effect of baseline  
200 shift. Secondary differentiation was applied also to the spectrum of magnesium  
201 stearate. The differentiated spectra of magnesium stearate and the methanol  
202 solutions of different magnesium stearate concentrations are shown in Figure 3.  
203 The magnesium stearate spectrum is scaled so that the spectral peaks can be  
204 clearly seen. More detailed information about the materials and experimental  
205 condition is described in Nakagawa et al. (Nakagawa et al., 2012).

206 In this case study, no scaling, autoscaling, and the proposed methods were  
207 compared. No scaling and autoscaling were applied with two popular statistical  
208 wavelength selection methods, *i.e.* VIP and LASSO. On the other hand, all  
209 wavelengths were used when the proposed methods were applied. From Table 1,  
210 the data from runs 1-9 was used for model building; 10-15 for parameter tuning;  
211 and 16-21 for model validation. To evaluate the influence of the number of  
212 samples on estimation accuracy, a different number of the model building and  
213 parameter tuning samples were used in cases 1 and 2. In case 1, one sample was  
214 randomly selected from each of runs 1-15, and 9 samples from runs 1-9 were for  
215 model building and 6 samples from runs 10-15 were used for parameter tuning.  
216 To evaluate the influence of sample selection on estimation performance, 100 sets  
217 of model building and parameter tuning data were independently generated. In  
218 case 2, all samples were used. Table 2 shows the model validation results. For  
219 case 1, the median, top 25<sup>th</sup> percentile (first quartile) and bottom 25<sup>th</sup> percentile  
220 (third quartile) of the RMSEs obtained from the 100 sets used for model building  
221 and parameter tuning data are shown. Tuning parameters such as the number  
222 of the latent variables in PLS models and the thresholds in VIP and LASSO were  
223 determined by trial and error so as to minimize the RMSE for the parameter tuning  
224 data. In proposed method 1 using VIP score, 5 latent variables were selected, and  
225 the iteration number  $i$  was determined as 5. The proposed methods gave 12-63%

226 smaller RMSE for model validation data than the conventional input variable  
227 scaling methods even when wavelength selection was conducted using VIP and  
228 LASSO. Figure 4 shows the VIP score for different number of iterations  $i$ . The  
229 VIP score with  $i = 1$  was used for wavelength selection in method 5, and that with  
230  $i = 5$  was used as input scaling factor in method 8. By the iterative calculation  
231 of the VIP score, important variables around 2800 and 1500 nm are emphasized,  
232 and the estimation performance was improved.

233 The above results clearly demonstrate the effectiveness of the proposed  
234 methods; even without variable selection they were able to reduce the estimation  
235 error. Proposed method 2 had about 10% smaller RMSE than proposed method  
236 1 in case 1, where the number of samples used for model building and parameter  
237 tuning is small. This result confirms that process knowledge is helpful for input  
238 variable scaling and can contribute to improve estimation performance.

Table 1: Experimental data for estimation of magnesium stearate concentration.

| Run number | Magnesium stearate<br>concentration [ $\mu\text{g}/\text{cm}^2$ ] | Number of samples |
|------------|---|-------------------|
| 1          | 0.08  | 5                 |
| 2          | 0.20  | 5                 |
| 3          | 0.40  | 5                 |
| 4          | 0.80  | 5                 |
| 5          | 1.20  | 5                 |
| 6          | 1.60  | 5                 |
| 7          | 2.88  | 5                 |
| 8          | 3.20  | 5                 |
| 9          | 4.00  | 5                 |
| 10         | 0.12  | 5                 |
| 11         | 0.24  | 5                 |
| 12         | 0.40  | 5                 |
| 13         | 0.80  | 5                 |
| 14         | 1.20  | 5                 |
| 15         | 1.60  | 5                 |
| 16         | 0.16  | 5                 |
| 17         | 0.32  | 5                 |
| 18         | 0.40  | 5                 |
| 19         | 0.80  | 5                 |
| 20         | 1.20  | 5                 |
| 21         | 1.60  | 5                 |

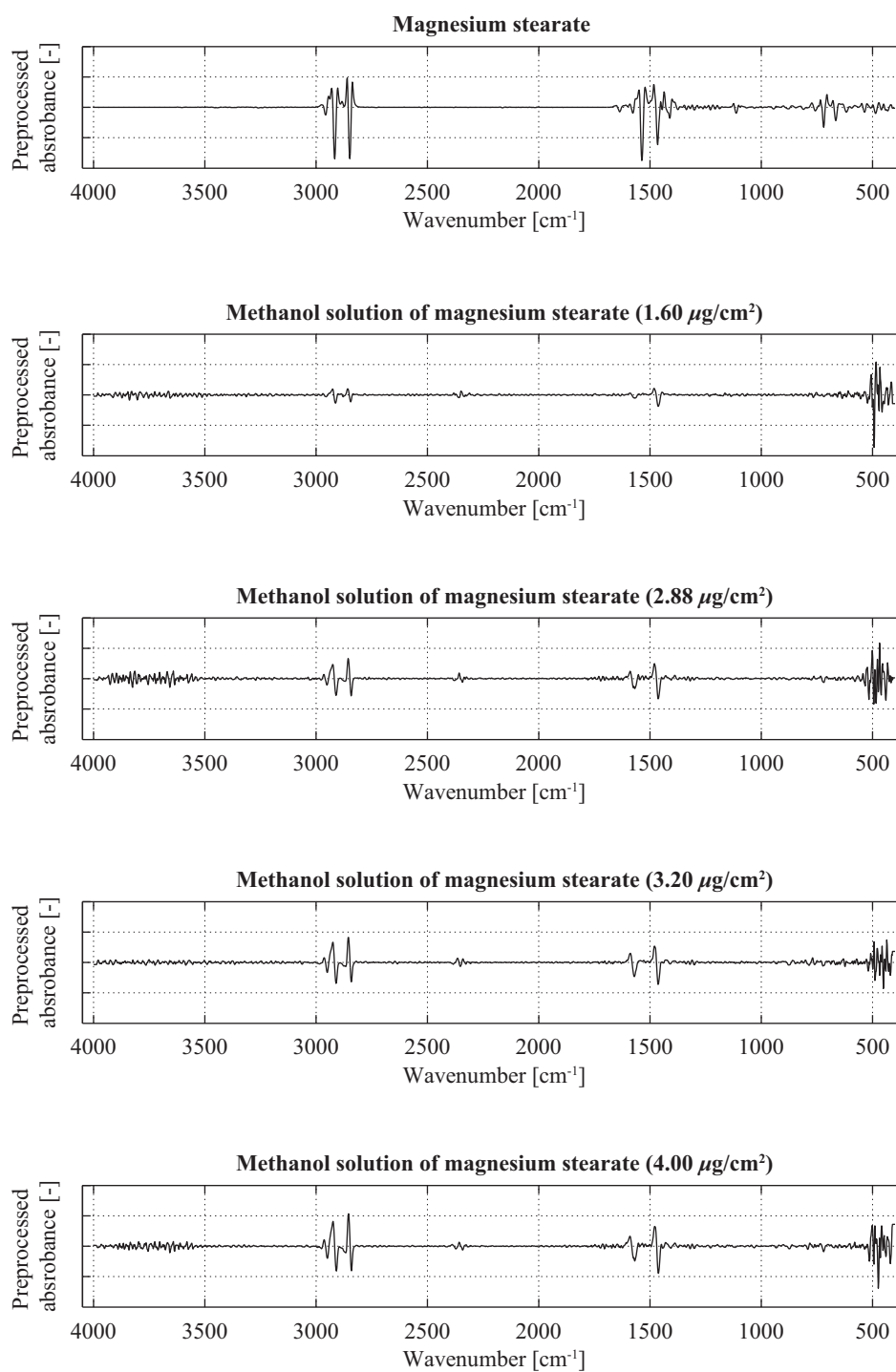


Figure 3: Spectra of magnesium stearate and methanol solutions at different magnesium stearate concentrations.

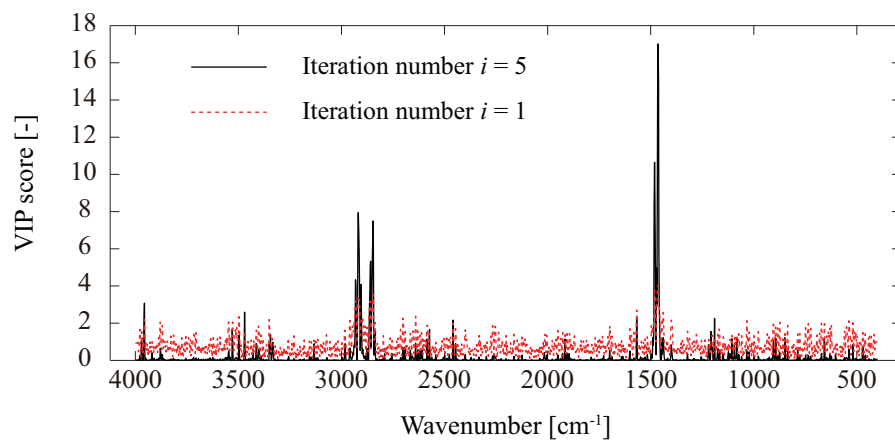


Figure 4: VIP score for the different iteration numbers.

Table 2: Results of the case study in the pharmaceutical process.

| Method | Scaling                        | Wavelength selection | Model | RMSE                  |        |
|--------|--------------------------------|----------------------|-------|-----------------------|--------|
|        |                                |                      |       | Case 1                | Case 2 |
| 1      | None                           | None                 | PLS   | 0.362 / 0.386 / 0.418 | 0.346  |
| 2      | None                           | VIP                  | PLS   | 0.363 / 0.386 / 0.419 | 0.346  |
| 3      | None                           | LASSO                | LASSO | 0.338 / 0.338 / 0.348 | 0.329  |
| 4      | Autoscaling                    | None                 | PLS   | 0.277 / 0.285 / 0.295 | 0.200  |
| 5      | Autoscaling                    | VIP                  | PLS   | 0.265 / 0.278 / 0.285 | 0.178  |
| 6      | Autoscaling                    | LASSO                | LASSO | 0.239 / 0.273 / 0.301 | 0.156  |
| 7      | Proposed method 1 (reg. coef.) | None                 | PLS   | 0.207 / 0.239 / 0.266 | 0.160  |
| 8      | Proposed method 1 (VIP)        | None                 | PLS   | 0.207 / 0.234 / 0.256 | 0.130  |
| 9      | Proposed method 2              | None                 | PLS   | 0.199 / 0.215 / 0.231 | 0.132  |

\*reg. coef.: regression coefficient



239 4.2. *Distillation process*

240 In distillation processes, soft sensors are often used to estimate product  
241 quality such as the concentration of impurities. Soft sensors were developed  
242 to estimate the 95% distillation temperature, which is an important quality of  
243 cracked gasoline. In the target process, the 95% distillation temperature is  
244 usually measured once a day, and a soft sensor is needed to implement inferential  
245 control of the 95% distillation temperature and to reduce the energy consumption.  
246 Forty-nine input variables, including 24 temperatures, 17 flow rates, 3 densities,  
247 2 pressures, and 3 liquid levels, were used for model building. Three hundred  
248 samples were used for model building. Data for parameter tuning and model  
249 validation both consisted of 100 samples. Tuning parameters such as the number  
250 of the latent variables in the PLS model and the thresholds for input variable  
251 selection were selected by trial and error so as to minimize the RMSE for the  
252 parameter tuning data.

253 Figure 5 shows the model validation results. In this example, autoscaling and  
254 proposed method 1 were compared. Proposed method 2 was not used since the  
255 spectrum of the product was not available. The values of the 95% distillation  
256 temperature were scaled so that the RMSE for model validation data of the  
257 conventional method using autoscaling without input variable selection was 1. As  
258 shown in Figure 5, proposed method 1 reduced the RMSE for model validation  
259 data by about 30% compared to the method using autoscaling without variable  
260 selection. As well, proposed method 1 using VIP scores reduced the RMSE by  
261 about 10% compared to methods using autoscaling with VIP and LASSO. This  
262 result confirmed the usefulness of proposed method 1.

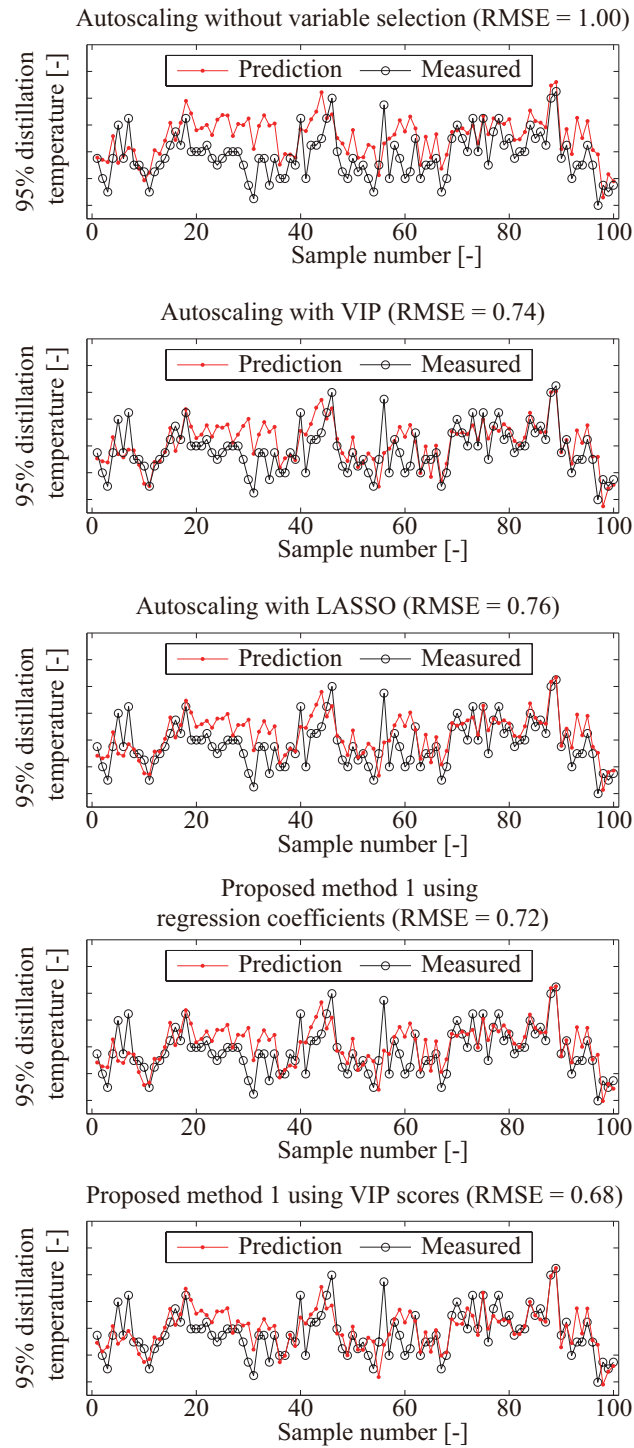


Figure 5: Model validation result in the distillation process.

263 **5. Conclusions**

264 This paper on input variable scaling methods for soft-sensor design showed  
265 that the input variable scaling factors should be determined on the basis of the  
266 importance of input variables for output estimation. Two new input variable  
267 scaling methods, which can evaluate the importance of input variables, were  
268 proposed. One method statistically derives the input variable scaling factors. The  
269 other one uses the spectroscopic data of a material whose content is an estimation  
270 target. The effectiveness of the proposed methods was confirmed through their  
271 application to a numerical example and industrial applications in a pharmaceutical  
272 and a distillation processes. The proposed methods were able to develop up to  
273 63% more accurate soft sensors compared to the conventional methods such as  
274 autoscaling with variable selection methods.

**References**

- Barnes, R.J., Dhanoa, M.S., Lister, S.J., 1989. Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. *Appl. Spectrosc.* 43, 772–777.
- van den Berg, R.A., Hoefsloot, H.C.J., Westerhuis, J.A., Smilde, A.K., van der Werf, M.J., 2006. Centering, scaling, and transformations improving the biological information content of metabolomics data. *BMC Genomics* 7:142.
- Cen, H., He, Y., 2007. Theory and application of near infrared reflectance spectroscopy in determination of food quality. *Trends Food Sci. Technol.* 18, 72 – 83.
- Engel, J., Gerretzen, J., Szymańska, E., Jansen, J.J., Downey, G., Blanchet, L., Buydens, L.M., 2013. Breaking with trends in pre-processing? *Trends in Anal. Chem.* 50, 96 – 106.
- Fujiwara, K., Sawada, H., Kano, M., 2012. Input variable selection for PLS modeling using nearest correlation spectral clustering. *Chemom. Intell. Lab. Syst.* 118, 109–119.
- Hocking, R.R., 1976. A biometrics invited paper. the analysis and selection of variables in linear regression. *Biometrics* 32, pp. 1–49.
- Jamrógiewicz, M., 2012. Application of the near-infrared spectroscopy in the pharmaceutical technology. *J. Pharmaceut. Biomed.* 66, 1 – 10.

- Jiang, J.H., James, R., Siesler, B.H.W., Ozaki, Y., 2002. Wavelength interval selection in multicomponent spectral analysis by moving window partial least-squares regression with applications to mid-infrared and near-infrared spectroscopic data. *Anal. Chem.* 74, 3555–3565.
- Jouen-Rimbauda, D., Massart, D.L., 1995. Genetic algorithms as a tool for wavelength selection in multivariate calibration. *Anal. Chem.* 67, 4295–4301.
- Kadlec, P., Gabrys, B., Strandt, S., 2009. Data-driven soft sensors in the process industry. *Comput. and Chem. Eng.* 33, 795–814.
- Kano, M., Fujiwara, K., 2013. Virtual sensing technology in process industries: trends and challenges revealed by recent industrial applications. *J. Chem. Eng. Jpn.* 46, 1–17.
- Keun, H.C., Ebbels, T.M.D., Antti, H., Bollard, M.E., Beckonert, O., Holmes, E., Lindon, J.C., Nicholson, J.K., 2003. Improved analysis of multivariate data by variable stability scaling application to nmr-based metabolic profiling. *Anal. Chim. Acta* 490, 265–276.
- Khatibisepehr, S., Huang, B., Khare, S., Domlan, E., Xu, F., Espejo, A., Kadali, R., 2014. A probabilistic framework for real-time performance assessment of inferential sensors. *Control Engineering Practice* 26, 136 – 150.
- Kim, S., Kano, M., Hasebe, S., Takinami, A., Seki, T., 2013. Long-term industrial applications of inferential control based on just-in-time soft-sensors: Economical impact and challenges. *Ind. Eng. Chem. Res.* 52, 12346–12356.
- Kim, S., Kano, M., Nakagawa, H., Hasebe, S., 2011. Estimation of active pharmaceutical ingredients content using locally weighted partial least squares and statistical wavelength selection. *Int. J. Pharm.* 421, 269–274.
- Kuzmanovski, I., Novi, M., Trpkovskaa, M., 2009. Automatic adjustment of the relative importance of different input variables for optimization of counter-propagation artificial neural networks. *Anal. Chim. Acta* 642, 142–147.
- Martens, H., Hoy, M., Wise, B.M., Bro, R., Brockhoff, P.B., 2003. Pre-whitening of data by covariance-weighted pre-processing. *J. Chemom.* 17, 153–165.
- Nakagawa, H., Tajima, T., Kano, M., Kim, S., Hasebe, S., Suzuki, T., Nakagami, H., 2012. Evaluation of infrared-reflection absorption spectroscopy

- measurement and locally weighted partial least-squares for rapid analysis of residual drug substances in cleaning processes. *Anal. Chem.* 84, 3820–3826.
- Nørgaard, L., Saudland, A., Wagner, J., Nielsen, J.P., Munck, L., Engelsen, S.B., 2000. Interval partial least-squares regression (iPLS): A comparative chemometric study with an example from near-infrared spectroscopy. *Appl. Spectrosc.* 54, 413–419.
- Oh, S.K., Yoo, S.J., Jeong, D.H., Lee, J.M., 2013. Real-time estimation of glucose concentration in algae cultivation system using raman spectroscopy. *Bioresource Technology* 142, 131 – 137.
- Rajalahti, T., Kvalheim, O.M., 2011. Multivariate data analysis in pharmaceuticals: A tutorial review. *Int. J. Pharm.* 417, 280–290.
- Roggo, Y., Chalus, P., Maurer, L., Lema-Martinez, C., Edmond, A., Jent, N., 2007. A review of near infrared spectroscopy and chemometrics in pharmaceutical technologies. *J. Pharm. Biomed. Anal.* 44, 683–700.
- Savitzky, A., Golay, M.J.E., 1964. Smoothing and differentiation of data by simplified least squares procedures. *Anal. Chem.* 36, 1627–1639.
- Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* 58, 267–288.
- Todeschini, R., Consonni, V., Maiocchi, A., 1999. The k correlation index theory development and its application in chemometrics. *Chemom. Intell. Lab. Syst.* 46, 13–29.
- Wold, S., Sjöström, M., Eriksson, L., 2001. PLS-regression: a basic tool of chemometrics. *Chemom. Intell. Lab. Syst.* 58, 109 – 130.